A Comparative Analysis of Different Models Explaining the Relationship Between Instructor Ratings and

Expected Student Grades

Robert E. Wright
University of Illinois

John C. Palmer
Quincy University

The widespread use of student evaluations to rate faculty has raised the question of whether high student evaluations can be gained simply through the process of faculty giving higher grades to students, or whether learning of students is a critical factor in such evaluations. Four different models were tested which represented different relationships between students= expected student grades and student evaluations of the quality of instructors, with and without student motivation, ability, and amount learned as potentially important variables. Evaluations from 119 students of four different instructors were used for the data set. Statistical tests of the alternative models indicated that a more complex model incorporating student motivation and ability levels as factors affecting student evaluations of instructors provided the best fit to the data. The fit was superior to that of a model using only expected grades and student evaluations of instructors, indicating that students= evaluations of faculty did not appear to be based solely on the grades students expected to receive. The complex model also fit the data better than a simpler model using only perceived amount learned, expected grades, and instructor ratings. For this data set, instructor ratings were not simply a function of expected grades, or simply a function of perceived amount learned, but a function of motivation, ability, amount learned, and grades.

In most colleges and universities in the U.S., students have long evaluated the performance of their instructors at the end of academic terms (e.g., Harrison, et al., 2004; Magner, 1997, Smith, 2004). Results of these evaluations are frequently utilized by individuals involved in personnel processes as a key criterion in making tenure and promotion decisions (Ehie & Karathanos, 1994; Harrison, et al., 2004; Smith, 2004; Williams & Ceci, 1997).

After extensive reviews of the literature, Marsh (1987) and Ellis, et al., (2004) found numerous studies that had reported positive relationships between the grades students expected to receive in classes and student ratings of instructors. Marsh (1987) and Ellis, et al. (2004) noted that, to the extent that this positive relationship may reflect grading leniency independent of other instructional attributes, such assessments might lack utility in measuring teaching effectiveness. However, Marsh and Ellis, et al. also noted that valid student evaluations could exhibit this same relationship, if, in fact, more effective teaching resulted in both higher expected grades and higher instructor ratings.

A number of other researchers have also examined this issue. For example, following an extensive examination of written comments on student evaluations, Trout (1997) specifically noted that level of course rigor appeared to be negatively associated with student ratings of instructors. Greenwald and Gillmore (1997) also concluded that ratings of instructors were affected by grading leniency, and described a statistical method that could be used to remove such contamination. Similarly, Ellis, et al. (2004) found evidence that the average student grade given in a course was a significant predictor of average student ratings of instructional quality of that course, and also suggested a need for adjusting student evaluations based on grades for a class. Krautman and Sander (1999) also found that high grades were related to higher teaching evaluations, and noted that such evidence indicated that such evaluations were a flawed measure of teaching performance. Similarly, McKeachie (1997) noted that care should be taken in how student ratings of instructors are utilized for personnel decisions, because, in some cases, higher grades may be given by instructors in an attempt to produce more positive teaching evaluations. Crumbley, et al., (2004) found that student evaluations might have encouraged a lack of rigor in the classroom on the part of instructors in order to gain higher evaluations. On a related note, Barry and Thompson (1997) and Marks (2000) also found that grading, specifically perceived fairness in grading, was an important predictor of overall student

assessments of faculty.

While many authors cite studies showing widespread evidence those higher expected grades are associated with higher student evaluations of faculty as clear evidence of a problem with the use of such evaluations, others disagree about the extent of such a problem. For example, Howard and Maxwell (1982) were unable to determine whether effective teaching caused high expected grades, and thus high levels of satisfaction for students, or whether high levels of student satisfaction were simply a function of higher expected grades. O=Connell and Dickinson (1993) found a high correlation between perceived learning and instructor ratings, which would indicate a high level of utility for instructor ratings. d'Appolonia and Abrami (1997) and Boretz (2004) pointed out that grading leniency is only a problem when it is unrelated to student learning. To the extent that student evaluations of instructors reflect student learning, such evaluations would be extremely useful in evaluation of instructors. However, to the extent that instructor ratings are simply a function of high grades given to students regardless of learning, the ratings would be of little use in evaluation of instructors.

Examining the relationship between expected student grades and student evaluations of instructors is of particular importance to educators and administrators. College graduates are typically expected by employers to have a basic understanding of a common body of knowledge in their field. To the extent that student evaluations of instructors reflect lenient grading policies of instructors, rather than fundamental knowledge gained by students, these evaluations would lack utility for use in assessing and rewarding instructor performance. Use of a flawed measure could result in rewarding faculty who may not be producing students capable of meeting the challenges of the world outside of the college or university. This could lead to reduced hiring of such students. However, if such evaluations were truly reflective of student learning of the key concepts taught in schools, the use of student evaluations would be an

effective tool in evaluating and rewarding faculty performance, and producing desired outcomes in terms of student learning.

The association of high student evaluations of instructors with high grades of students may result from students rewarding lenient instructors, or from students learning a great deal (and thus obtaining high grades), and in turn, rewarding the teaching which led to the high level of learning. However, neither explanation may adequately explain the positive relationship between student grades and instructor ratings. Rather, a more complex explanation may be needed. Marsh (1987) suggested that, in addition to the possibility that grading leniency or validity of student evaluations explained the relationship between student grades and instructor ratings, a more complex relationship might exist, with factors such as student ability and motivation being important predictors of instructor ratings.

If initial differences in characteristics of students such as motivation and ability affect both student performance (in terms of both learning and grade expected in classes) and student perceptions of teaching effectiveness, evaluations of instructors by students may be difficult to interpret.

If the students are not very motivated and/or have low levels of ability, they may prefer a lenient instructor who does not require that students learn a great deal. They may reward such leniency, while punishing a rigorous instructor, who requires a large amount of learning in order for students to be successful, for the students= lack of success in the class. However, if students are highly motivated and have a high level of ability, they may prefer an instructor who challenges them to learn and understand material at a high level, and reward such an instructor with higher ratings. Conversely, these students may punish an instructor who requires little learning of the subject matter in order for students to succeed in the classroom.

Purpose

This purpose to this study was to investigate the utility of student ratings of instructors by comparing different models of the

relationships among motivation, ability, perceived amount learned, expected grades, and ratings of instructors. For student evaluations of instructors to be useful, they should be strongly reflective of student learning, based on the amount students learned of material they should have learned in the class. If the evaluations strictly reflect leniency in grading, they would have little value.

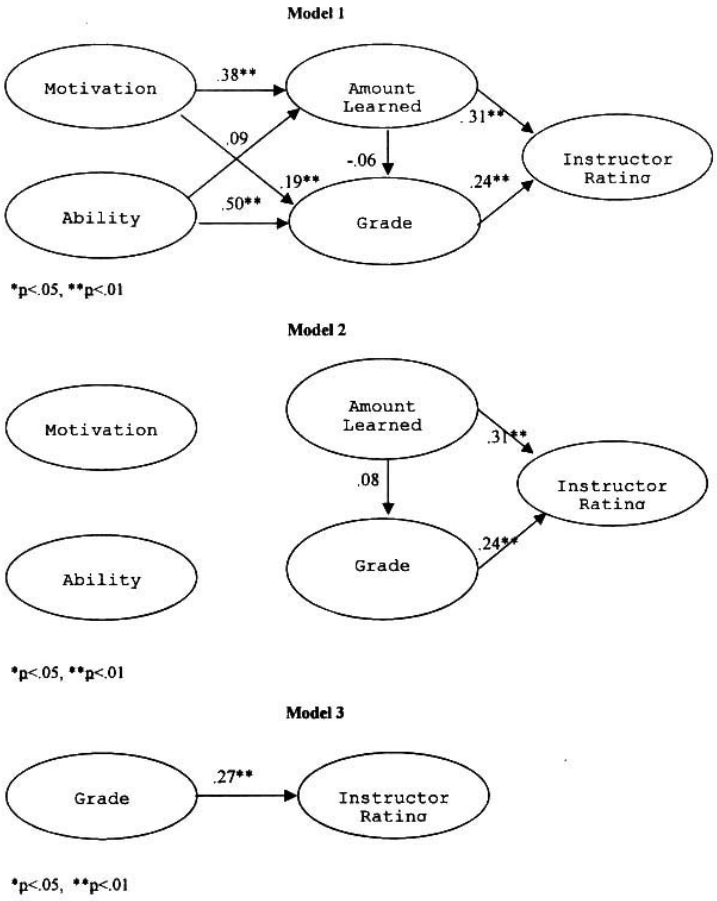
One potential difficulty in studying this topic was the ability of students to rate what they should have learned. Marks (2000) raised the question of how students can rate whether they have learned what they should have learned without expertise in the field. However, students should have the ability to rate their own perceptions of the amount they learned. This study will therefore use student perceptions of the amount learned in order to explore relationships between expected student grades and instructor ratings. This sidesteps the difficulty of dealing with the problem of whether students are able to determine the actual amount they learned by using perceived amount learned as a proxy for the actual amount learned.

Student ratings of instructors are used to measure student perceptions of the class experience. If students do, in fact, rate instructors based mainly on the students= perceived amounts of learning, the student evaluations would be worthwhile measures of at least one aspect of the instructor=s performance. However, if the instructor ratings were a result of other factors in addition to perceived amount learned, this could cause great difficulty in interpreting the student ratings for use in evaluating instructors.

Method

To evaluate possible explanations for the relationship between expected grades and instructor ratings, three models were developed that would approximate different explanations for such a relationship.

Figure 1: Models 1, 2, and 3



The complex model (Figure 1, Model 1) posits that motivation and ability affect both the perceived amount learned and the expected grade, which in turn affect the instructor rating.

In the optimal case, student evaluations of instructors would be based on the perceived amount learned by students,

which would be reflected in the grades students expected to receive. Figure 1 (Model 2) reflects this situation, with a path from perceived amount learned to both grade and instructor rating, and a path from grade to instructor rating. In this situation, the relationship between expected grades and instructor ratings is because the amount learned is reflected in the grade, as well as in the instructor rating.

The simple model would reflect the situation where instructor ratings were simply a function of students= expected grades. Figure 1 also shows this model (Model 3). The only path in this model is one from grades to instructor rating. All other paths in the model are set to zero, indicating no other variables affect this relationship.

To test these models, student surveys from classes taught by four different instructors at a medium sized Midwestern university were gathered. All students surveyed (a total of 119) completed the surveys.

Students indicated expected grades by circling the appropriate letter grade between A and F, with pluses and minuses included. The student evaluations were given out the last two weeks of class, so students would have a good idea of what grade to expect. Letter grades were converted to a numeric value, with A equal to 1, A- equal to 2, etc. The average grade was 2.65 (approximately a B+), with a standard deviation of 1.54.

Perceived amount learned was measured using a five point scale, ranging from "A large amount" (recorded as a 1) to "Very Little" (recorded as a 5). Average level of perceived learning was 1.45, with a standard deviation of .582. In order to reduce problems with instrumentation differences affecting possible relationships, instructor ratings were measured on a scale similar to the student expected grade scale. Instructor rating was measured on a 5 point scale, with 1 being outstanding, and 5 being poor. The average instructor rating was a 2.22, with a standard deviation of 1.34. Motivation was measured on a 5-point scale, with 1 being "extremely motivated" and 5 being

"not at all motivated." The average rating was 1.95, with a standard deviation of 0.72. Ability was a self reported measure, on a nine-point scale, from 1 being "top 1% of the students at the university" to 9 being "bottom 25% of the students at the university." The mean was 4.10, with a standard deviation of 1.91. The low mean and wide spread in responses to this question indicated that students appeared to attempt to answer this question truthfully.

Table 1: Intercorrelations Between Variables

	Amt. Learned	Grade	Motivat.	Instr. Rating	Ability
Amt. Learned	1.00	-.02	0.40**	0.34**	0.08
Grade		1.00	0.20*	0.25**	0.51**
Motivat			1.00	0.31**	0.05
Instr. Rating				1.00	0.03
Ability					1.00

*p<.05, **p<.01

Initially, correlations between variables were observed (see Table 1) to determine if instructors= ratings were positively correlated with students= expected grades. The correlation between the instructor rating and students= expected grades was .25 (p<.01).

Once the positive correlation between students= expected grades and instructor ratings was established, the next step was to test the models representing different relationships among variables.

In order to test these competing models with multiple, interdependent relationships, the models previously described were created using LISREL 8 (Joreskog & Sorbom, 1996). This technique provided for analysis of the significance of both the paths in the models, and the overall fit of the model to the data. Using this type of structural equations modeling therefore allowed more complex models to be tested which reflected different possible explanations of the relationship between instructor ratings and expected student grades. Models were developed with paths between variables free to be estimated, or with certain paths constrained to zero. The models were then compared to determine which model best fit the observed data.

Data was analyzed using LISREL 8 (Joreskog & Sorbom, 1996). The sample size was within the 100 to 200 range suggested as optimal for testing structural equations models by Hair, Jr., Anderson, Tatham, and Black (1995, p. 637). Three nested models (e.g., Bentler & Bonnet, 1980) were initially compared to determine which best represented the data. (Nested models were created by using the same variables for all models, and then restricting an increasing number of paths among the variables for each succeeding model by setting certain paths equal to zero.) The variables used in all the models were motivation, ability, perceived amount learned, expected grade in the class, and overall rating for the instructor.

Results

The fit of each model was evaluated with commonly used measures of fit including the Chi-square goodness of fit test (Bentler & Bonnet, 1980), and the Goodness of Fit index and Normed Fit Index (Bentler, 1990.)

Model 1 (see Figure 1) produced a good statistical fit to the data (Chi-square (2 d.f.) = 4.65, $p > .05$). (This result

indicates that the hypothesis that the data reflects the proposed model cannot be rejected. Conversely, *p* values less than .05 would indicate that the hypothesis that the data reflects the proposed model should be rejected at the .05 level of statistical significance.) In addition, the Goodness of Fit (GFI) Index was .98, and the Normed Fit Index (NFI) was .95. The number of degrees of freedom was also close to the Chi square value, which also indicates a good fit.

Model 2 (see Figure 1), was not a good statistical fit (Chi square (6 d.f.) = 62.56, $p < .01$). In addition, the GFI was .84, and the NFI was .26, also indicating a poor fit. Bentler and Bonnet (1980) noted, "models with overall fit indices of less than .90 can usually be improved substantially" (*p*.600.)

Model 3 (see Figure 1), with only one path from expected student grades to instructor ratings, was not a good statistical fit (Chi square (8 d.f.) = 75.73, $p < .00$). In addition, the GFI of .80, and the NFI of .11, also indicate a poor fit.

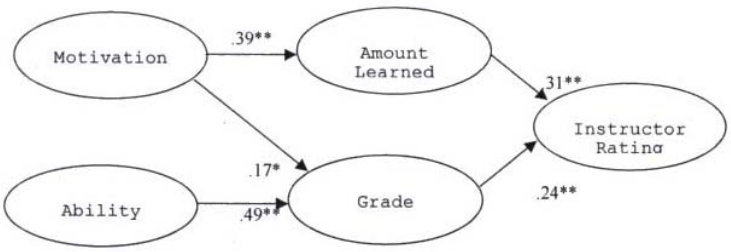
Because the models are nested, a comparison of the change in the value of Chi-square relative to the change in degrees of freedom can also be used to compare the models (Bentler & Bonnet, 1980).

The change in Chi-square from the Model 2 to Model 1 was 58.09, with a change in degrees of freedom of 4. This is statistically significant at the .01 level. The change in Chi-square from Model 3 to Model 1 was 71.08 with 6 degrees of freedom, which is also statistically significant at the .01 level. This comparison shows Model 1 to provide a statistically significantly better fit to the data than the other models.

However, the Model 1 analysis showed that the path from ability to amount learned was non-significant, and the path from amount learned to grade was also non-significant. Re-estimating the model with those two paths constrained to be equal to zero gave a Chi square statistic of 6.13, with 4 degrees of freedom ($p > .15$). Thus, this modified Model 1 (see Figure 2)

Figure 2

Adjusted Model 1



provided a good statistical fit to the data. The model had a GFI of .98, and a NFI of .93, both indicating a very good fit to the data. While the improvement in Chi square over the Model 1 was only 1.51, with 2 degrees of freedom, which is not a statistically significant improvement from the other tests, as well as the correlation matrix (see Table 1) suggests the model eliminating the paths from ability to amount learned, and from amount learned to grade is more reflective of the data.

Data analysis showed that Model 1 fit the data much better than either Model 2 or Model 3. The modified version of Model 1 proved to be the best fitting model for the data set, based on commonly used fit indices.

Discussion

Of the three original models developed based on the literature, the model which included motivation and ability as positive predictors of perceived amount learned and expected grade, and perceived amount learned and expected grade as positive predictors of instructor rating, provided the best fit to the data. However, a modified version of this model fit the sample data better than the original model. Evidence from this sample suggests that neither simple explanation for the relationship between expected student grades and instructor ratings accurately fit the data. Instructor ratings did not appear to be solely a function of the grade students expected to receive,

nor did they appear to be strictly a function of the perceived amount learned by students.

If instructor ratings had been simply a function of the grades students expected to receive, all other paths in the model would have been irrelevant. However, results of the data analysis clearly indicate that a model using only one path, from expected grades to instructor ratings, does not fit the data well. In this data set, it appears that being a lenient instructor does not, by itself, result in high ratings. This finding casts some doubt on those who claim that instructors may get high teaching evaluations solely from being easy graders.

However, neither did the data support the idea that instructor ratings were solely the result of the perceived amount learned. The model representing this explanation also did not fit the data well. Therefore, it appears that the data do not support those authors suggesting that student evaluations are a very effective measure of teaching.

Rather, the relationship between expected student grades and instructor ratings appeared to be much more complex. Instructor ratings appeared to be a function of both the perceived amount learned and the grade expected by the students, with both higher perceived amounts learned and higher grades leading to higher instructor ratings. Further, higher perceived amounts learned were a function of higher motivation levels of the student, and the expected grade earned was a function of both the motivation and ability level of the student, with higher levels of motivation and ability leading to higher expected grades.

These finding indicate that student evaluations of instructors may be a good measure of performance for students whose motivation and ability are in line with the instructor=s expectations. However, for those students whose motivation and ability levels are not in line with instructor expectations, the instructor evaluations may not be a good measure of instructor performance.

Somewhat unexpectedly, perceived amount learned was not a function of the ability level of the students in the study. However, a possible explanation for this finding could be that

students of lower ability in the sample might have worked hard and learned a great deal, while students of higher ability may not have worked as hard, learning much less.

Another unexpected finding from the study was the lack of relationship between perceived amount learned and expected grade. However, this result may be due to the effects of ability on grade. Some students in the sample may have worked very hard and learned a great deal. However, due to a lower level of ability, they still obtained a relatively low grade. Other students may not have worked as hard, and thus not learned a great deal, but due to a high level of innate ability obtained high grades.

Conclusion

These results point out the possible problems with the use of student evaluations of teaching as the only instrument for measuring teaching effectiveness. Even while using students= perceived amounts of learning, rather than actual student learning, as a criterion of teaching effectiveness, this research showed that evaluations may be a useful measure of teaching effectiveness only for a specific group of students in a given class who have particular levels of motivation and ability that are in line with the instructor's expectations for the course.

Students in the sample did not seem to reward professors with high teaching evaluations solely because the students believed that they would receive high grades. However, students also did not seem to reward professors with high teaching evaluations simply because the students believed that they learned a great deal.

The relationship between grades and teaching evaluations appeared to be due to a more complex relationship among perceived amount learned, motivation, and ability.

Results from this study would indicate that it might be extremely difficult to interpret an individual instructor's ratings, due to the possible wide dispersion of motivation and ability levels in a particular class. Comparisons across subject matter and academic disciplines also might be quite difficult. While

certain classes might have students with similar levels of motivation and ability, other classes might have students with a wide variety of motivation and ability levels.

Results from this study may aid in explaining the variety of results that have been obtained when researchers have attempted to assess the validity of student evaluations of instructors. While some studies have provided evidence that perceived student learning is a key factor in instructor evaluations, others have found that such evaluations seem dependent on factors that may be unrelated to learning, such as instructor leniency. Whether or not student evaluations of instructors are related to perceived student learning might depend on whether instructor expectations are consistent with the motivation and ability levels of students.

Given the importance of students learning as much as possible in order to increase their chances of future success, as well as the increasing importance of student evaluations on promotion and tenure decisions concerning faculty at colleges and universities, knowledge of the effects of such factors as student grades, student learning, student motivation, and student ability on faculty evaluations is vital. Future research should further examine such relationships in a variety of situations, with a variety of teachers and subjects, in order to determine how variations in such factors affect these relationships.

Administrators in higher education using student evaluations of instructors as a key indicator of instructor effectiveness should be aware that interpreting results of such measures might be problematic.

References

Barry, T. & Thompson, R. (1997). Some intriguing relationships in business teaching evaluations. *Journal of Education for Business, 72*, 303-311.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.

Bentler, P.M., & Bonnet, D.G. (1980). Significance tests and goodness of fit in analysis of covariance structures.

Psychological Bulletin, 88, (3), 588-606.

Boretz, E. (2004) Grade inflation and the myth of student consumerism. *College Teaching, 52*, 42-47.

Crumbley, L. Henry, B.K., & Kratchman, S.H. (2001). Student perceptions of the evaluation of college teaching. *Quality Assurance in Education, 9*, 197-207.

d'Apollonia, S. & Abrami, P. (1997). Navigating student ratings of instruction. *American Psychologist, 52*, 1198-1208.

Ellis, L., Burke, D.M., Lomire, P., & McCormack, D.R. (2004). Student grades and average ratings of instructional quality: The need for adjustment. *Journal of Educational Research, 97*, 35-41.

Ehie, I. C., & Karathanos, D. (1994). Business faculty performance evaluation based on the new AACSB accreditation standards. *Journal of Education for Business, 69*, (5), 257-262.

Greenwald, A. & Gillmore, G. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*, 1209-1217.

Hair, Jr., J., Anderson, R.E., Tatham, R.L., & Black, W.C. (1995). *Multivariate Data Analysis*, Fourth Edition, Prentice-Hall, Englewood Cliffs, N.J.

Harrison, P.D., Douglas, D.K., & Burdsal, C.A. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education, 45*, 311-323.

Howard, G. S. & Maxwell, S.E. (1982). Do grades contaminate student evaluations of instruction? *Research in Higher Education, 16*, 175-187.

Joreskog, K. & Sorbom, D. (1996). LISREL 8: *User's reference guide*. Scientific Software International, Inc., Chicago, Il.

Krautman, A.C. & Sander, W. (1999). Grades and student evaluations of teachers. *Economics of Education Review, 18*, 59-63.

Marks, R.B. (2000). Determinants of student evaluations of

global measures of instructor and course value. *Journal of Marketing Education, 22*, 108-119.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, (3), 253-388.

Magner, D. K., (1997) Report says standards used to evaluate research should also be used for teaching and service. *The Chronicle of Higher Education, 44*, (2), A18-A19.

McKeachie, W. (1997). Student ratings: The validity of use. *American Psychologist, 52*, 1219-1225.

O'Connell, D. Q. & Dickinson, D.J. (1993). Student ratings of instruction as a function of testing conditions and perceptions of amount learned. *Journal of Research and Development in Education, 27*, (1), 18-23.

Smith, G. S. (2004). Assessment strategies: What is being measured in student course evaluations? *Accounting Education, 13*, 3-28.

Trout, P. A. (1997). What the numbers mean: Providing a context for numerical student evaluations of courses. *Change, 29*, 25-30.

Williams, W.M. & Ceci, S.J. (1997). "How'm I doing?" Problems with student ratings of instructors and courses. *Change, 29*, 13-23.